*Statistical Applications in Genetics and Molecular Biology*

*Volume* 11, *Issue* 5 2012 *Article* 1

# DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms

**David M. Ramsey,** *University of Limerick*
**Andreas Futschik,** *University of Vienna*

**Recommended Citation:**
Ramsey, David M. and Futschik, Andreas (2012) "DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms," *Statistical Applications in Genetics and Molecular Biology*: Vol. 11: Iss. 5, Article 1.

# DNA Pooling and Statistical Tests for the Detection of Single Nucleotide Polymorphisms

David M. Ramsey and Andreas Futschik

## Abstract

The development of next generation genome sequencers gives the opportunity of learning more about the genetic make-up of human and other populations. One important question involves the location of sites at which variation occurs within a population. Our focus will be on the detection of rare variants. Such variants will often not be present in smaller samples and are hard to distinguish from sequencing errors in larger samples. This is particularly true for pooled samples which are often used as part of a cost saving strategy. The focus of this article is on experiments that involve DNA pooling. We derive experimental designs that optimize the power of statistical tests for detecting single nucleotide polymorphisms (SNPs, sites at which there is variation within a population). We also present a new simple test that calls a SNP, if the maximum number of reads of a prospective variant across lanes exceeds a certain threshold. The value of this threshold is defined according to the number of available lanes, the parameters of the genome sequencer and a specified probability of accepting that there is variation at a site when no variation is present. On the basis of this test, we derive pool sizes which are optimal for the detection of rare variants. This test is compared with a likelihood ratio test, which takes into account the number of reads of a prospective variant from all the lanes. It is shown that the threshold based rule achieves a comparable power to this likelihood ratio test and may well be a useful tool in determining near optimal pool sizes for the detection of rare alleles in practical applications.

**KEYWORDS:** genome sequencing, optimal DNA pooling, statistical inference, single nucleotide polymorphism

**Author Notes:** David Ramsey acknowledges the support of Science Foundation Ireland under the BIO-SI pro ject [grant no. 07MI012].

# 1 Introduction

The field of genome sequencing requires new statistical methods due to the introduction of new generation genome sequencers and the mass of data accompanying it. A genome may be thought of as a sequence where each site can be occupied by one of four nucleotides, denoted A, C, G and T. At a large majority of sites, each individual in a population has the same nucleotide. Sites at which a population shows variation are called single nucleotide polymorphisms (SNPs). In general, there are two variants (alleles) at such sites. The least (most) common of these two alleles is called the minor allele (major allele, respectively).

As it is simple to detect a SNP where both alleles are common, we concentrate on the detection of rare minor alleles. This problem is not trivial, since it is necessary to distinguish a low number of reads from individuals with a rare allele from sequencing errors. Compared to separate sequencing of individuals, pooling is a cost effective sequencing strategy, where DNA material from more than one individual is placed in a single lane of the sequencer. Large pools increase the chance of capturing rare alleles, but make it more difficult to avoid false positives due to sequencing errors. It thus makes sense to look for a pool size that is optimal in terms of power, given a specified probability of a type I error. We introduce two statistical tests for identifying SNPs based on pooled samples. One is a likelihood ratio test, whereas the other uses the maximum number of reads of the minor allele across all lanes. This maximum test is very simple and has a power comparable to the likelihood ratio test. We derive asymptotically optimal pool sizes for detecting rare alleles (i.e., pool sizes that are optimal when the frequency of a rare allele is small) according to the number of lanes available and the parameters of the sequencer used. Also, we investigate the minimum number of reads of an allele required from a lane to infer that a minor allele is present.

Accurate detection of SNPs is important in determining the genetic factors behind certain diseases, as well as the estimation of mutation rates. In population genetics, a local excess of SNPs where the minor allele is rare is a typical characteristic of genomic regions that have undergone recent positive selection. Also, SNP databases such as dbSNP (http://www.ncbi.nlm.nih.gov/snp) are accepting submissions of newly identified SNPs for many different organisms. Since there should be confidence in the SNPs submitted, sufficient statistical evidence is desirable, suggesting that an SNP found is not merely a sequencing error.

Achaz (2008) notes that interpreting sequencing errors as reads of a rare allele leads to a high false discovery rate when detecting SNPs. He proposes that the existence of a SNP should be inferred only when the number of reads of the minor allele exceeds a given threshold. Knudsen and Miyamoto (2009) and Jiang, Tavaré and Marjoram (2009) extend this work by proposing methods for estimating

the mutation rate (in effect estimating the number of SNPs along a section of the genome at the same time) taking sequencing errors into account. Jiang, Tavaré and Marjoram (2009) also propose a threshold rule for inferring genotypes.

When DNA is not pooled, genetic material from one individual is placed in each lane of a sequencer, which gives a random number of reads of the nucleotide (or the two nucleotides in the diploid case) at a particular site. Due to advances in sequencing, the mean number of reads from a site may be large (particularly when a section of the genome is amplified). In such cases, it may be more efficient to pool DNA (see Sham *et al.* (2002)). Although sequencing errors are more difficult to eliminate when sequencing pooled samples, the larger number of individuals that can be sequenced within a given budget will often outweigh these difficulties. Therefore, pooling is popular in practice. Holt *et al.* (2009) use pooling to estimate allele frequencies and detect SNPs in clonal bacterial populations. Druley *et al.* (2009) use a similar approach to finding rare alleles (i.e., detecting SNPs). Reads are anonymous in the sense that we do not know from whom a read came, only the lane. Hence, we cannot infer the genotypes of diploid individuals using such a pooling procedure. Craig *et al.* (2008) describe a method of "bar-coding", enabling researchers to infer which individual a read came from and use this approach to detect SNPs. However, bar coding is often impractical, particularly when genotyping small organisms. Kenny *et al.* (2011) use a similar approach to detecting SNPs. Also, they show that the distribution of the number of reads from individuals in a pool is relatively uniform. These methods use independent pools, i.e., each individual appears in exactly one pool. Hence, if there are $k$ lanes and the pool size is $m$, then the sample size is $km$. Erlich *et al.* (2009) use a pooling scheme in which individuals appear in several pools. Such a scheme enables us to infer which individual has an allele which is sufficiently rare in the sample without having to use bar-codes.

This article extends the results of Futschik and Schlötterer (2010) on the sequencing of anonymous, independently pooled DNA samples. It is assumed that the number of reads from a lane has a Poisson distribution and the probability of an error in any given read is a known constant. They consider the probability of detecting a SNP using a threshold rule, i.e., an allele is inferred to exist if the number of times it is read in a lane is greater than some chosen number. They also consider the probability of wrongly inferring that a site is a SNP. This article presents two tests for detecting SNPs which control the probability of wrongly inferring that a site is a SNP. One test is based on a threshold rule, which is analyzed to obtain pool sizes maximizing the probability of detecting a rare minor allele. It is assumed that there is a given number of available lanes and we have good estimates of the mean number of reads per lane and the error rate. Error rates can be obtained from PHRED-quality scores that are given for each read. Although these scores can

be translated into the probability of a sequencing error, these probabilities are not always very accurate. This is especially true for Illumina reads. Therefore, methods and software are available for recalibrating the scores. As these recalibrated scores may still not always be perfect, we also investigate the robustness of the results when estimates of the error probabilities are inaccurate and/or only an upper bound on the error probabilities is known.

Section 2 presents the problem and a simple statistical model describing the data obtained from genome sequencing. Section 3 presents a likelihood ratio test for detecting SNPs based on this statistical model. Section 4 presents a simple test based on the maximum number of reads of a prospective minor allele from a lane. A function estimating the power of such a test according to the pool size and the parameters of the genome sequencer is derived. The asymptotically optimal pool size is determined on the basis of this function. Section 5 gives results for simulations of the test procedures. In order to check the robustness of these tests, simulations were also carried out under more general assumptions. Section 6 summarizes the results and proposes directions for future research.

## 2 Formulation of the Problem

This section presents a simple model for the detection of SNPs. Assume that there are always two alleles at a SNP. Let the number of available lanes be $k$ and the number of individuals (assumed to be haplotypes) in a pool be $m$, i.e., the sample size is $n = km$. Assume that the number of reads of a site from a lane has a Poisson($\lambda$) distribution. The data are simply the number of reads from each lane of each nucleotide at each site along a section of the genome. It should be noted that for convenience $\lambda$ is assumed to be independent of the site and allele. Also, $\lambda$ is reasonably large (of order 10 or greater). Each read records an incorrect base with probability $\varepsilon$, independently of other reads (assumed to be of order $10^{-2}$ or lower). Note that we can also define $\varepsilon$ to be an upper bound on the error probabilities. In addition, suppose good estimates of the parameters $\lambda$ and $\varepsilon$ are available for the sequencer used.

Suppose that the minor allele frequency at a given locus is $p$. The goal is to define an optimal pooling procedure while controlling the type I error rate for the following hypotheses:

$\mathbf{H_0}$: The locus is not a SNP, i.e., $p = 0$.

$\mathbf{H_A}$: The locus is a SNP with minor allele frequency $p > 0$.

Given a reasonably large sample size, any sensible test will detect a minor allele of large frequency with power close to one. Hence, we concentrate on finding rare minor alleles. The argument for the existence of an optimal pool size based on

a test inferring the existence of a SNP when the maximum number of reads of the minor allele from a lane is above a certain threshold is as follows:

**1**. When the number of lanes is fixed, by pooling individuals we can increase the sample size. This increases the probability that a rare allele is actually included in the sample.

**2**. However, as the pool size increases, the number of reads per individual decreases and it becomes harder to distinguish between reads from one individual with the minor allele and errors.

**3**. Due to these counteracting effects, given that the number of reads per lane is reasonably large, the probability of detecting a rare minor allele has a maximum at some intermediate pool size.

Initially, we consider a model in which there are at most two different alleles at a site. Without sequencing errors, this corresponds to the infinite sites model (Balding, Bishop and Cannings , 2008). Assume that sequencing errors do not introduce a third base. This simplifies the real situation of three possible wrong choices for any nucleotide, but will be generalized for our simulations. The major allele is defined to be the allele with the largest number of reads in the sample as a whole. As we are interested in loci where the frequency of the minor allele is small, we can reasonably neglect the possibility that the major allele is not correctly identified. The other allele is the prospective minor allele.

Consider a given site. Let $R_i$ be the total number of reads for that site in lane $i$ and $\mathbf{R}=(R_1,R_2,\ldots,R_k)$ the vector of the number of reads for all $k$ lanes. By assumption the $R_i$ are independent and identically Poisson$(\lambda)$ distributed. Let $X_i$ be the number of reads of the prospective minor allele from lane $i$ and $\mathbf{X}=(X_1,X_2,\ldots,X_k)$. Define $A_i$ to be the number of individuals with the minor allele in lane $i$ and $\mathbf{A}=(A_1,A_2,\ldots,A_k)$. Note that $A_i \sim \text{Bin}(m,p)$ for a pool of size $m$. Let $B$ be the total number of individuals with the minor allele, i.e., $B=\sum_{i=1}^{k} A_i$. We observe $\mathbf{X}$ and $\mathbf{R}$.

Suppose that there is no-one with the minor allele in a pool. Hence, the number of reads of the prospective minor allele is simply the number of errors in that lane. Given $R_i=r$, $X_i$ has a binomial$(r,\varepsilon)$ distribution. Since $R_i$ has a Poisson distribution with expected coverage $\lambda$, $X_i$ has a Poisson$(\lambda\varepsilon)$ distribution. Now suppose that there are $a$ individuals with the minor allele in a pool, $a\in\{1,2,\ldots,m\}$. Assume that each individual contributes a large, equal amount of DNA. Hence, reads are obtained by binomial sampling from the pool. The empirical results of Kenny *et al.* (2011) show that this is a reasonable assumption. The distribution of $X_i$ given $R_i=r$ and $A_i=a$ is the binomial$(r,q(a;\varepsilon))$ distribution, where $q(a;\varepsilon)=\frac{a(1-\varepsilon)}{m}+\frac{\varepsilon(m-a)}{m}$. Hence, $X_i$ has a Poisson$(\lambda q(a;\varepsilon))$ distribution. Note that if $\varepsilon$ is small in comparison to $1/m$, then for $a\geq 1$, $q(a;\varepsilon)\approx a/m$.

# 3 A Likelihood Ratio Test

In this section, we derive a likelihood ratio test (LR test) of the null hypothesis that the reads of the prospective minor allele are purely due to sequencing errors. We treat the number of reads from a lane as an ancillary statistic (Lehmann , 1986, Chapter 10, Section 2). Hence, we compute our likelihoods conditional on $R_1,\ldots,R_k$. Under the alternative hypothesis, the likelihood function is also conditional on the (unobserved) number of individuals with the minor allele in each lane. Firstly, we consider the likelihood of the number of reads of a prospective minor allele from a single lane. Under the null hypothesis, we have

$$L_0(x_i|R_i=r_i)=\binom{r_i}{x_i}\varepsilon^{x_i}(1-\varepsilon)^{r_i-x_i}. \tag{1}$$

Similarly, under the alternative hypothesis and conditional on $A_i$

$$L_1(x_i|A_i=a_i,R_i=r_i)=\binom{r_i}{x_i}[q(a_i;\varepsilon)]^{x_i}[1-q(a_i;\varepsilon)]^{r_i-x_i}, \tag{2}$$

where $A_i\sim \mathrm{Bin}(m,p)$. In the case $A_i=0$, this is simply the likelihood under the null hypothesis.

Multiplying the null likelihoods for each lane, the denominator of the likelihood ratio is

$$\prod_{i=1}^{k}L_0(x_i|R_i=r_i)=\varepsilon^{\sum_{i=1}^{k}x_i}(1-\varepsilon)^{\sum_{i=1}^{k}(r_i-x_i)}\prod_{i=1}^{k}\binom{r_i}{x_i}. \tag{3}$$

Assume that the pools are obtained by independent binomial sampling from a large population where the minor allele frequency is $p$, the numerator of the likelihood ratio is given by

$$\begin{aligned}\prod_{i=1}^{k}L_1(x_i|R_i=r_i)&=\prod_{i=1}^{k}\sum_{a_i=0}^{m}L_1(x_i|A_i=a_i,R_i=r_i)P(A_i=a_i)\\&=\prod_{i=1}^{k}\sum_{a_i=0}^{m}\binom{m}{a_i}p^{a_i}(1-p)^{(m-a_i)}\binom{r_i}{x_i}[q(a_i;\varepsilon)]^{x_i}[1-q(a_i;\varepsilon)]^{r_i-x_i}.\end{aligned} \tag{4}$$

Setting $p=0$, we obtain the likelihood under the null hypothesis.

We now maximize the numerator with respect to the parameter $p$ and compute the likelihood ratio. This leads to

$$T=\frac{\max_p\prod_{i=1}^{k}\sum_{a_i=0}^{m}\binom{m}{a_i}p^{a_i}(1-p)^{(m-a_i)}[q(a_i;\varepsilon)]^{x_i}[1-q(a_i;\varepsilon)]^{r_i-x_i}}{\varepsilon^{\sum_{i=1}^{k}x_i}(1-\varepsilon)^{\sum_{i=1}^{k}(r_i-x_i)}}. \tag{5}$$

By taking logarithms, we get the LR test statistic $2\log(T)$.

Under suitable regularity conditions, the asymptotic properties of likelihood ratio tests are well understood. In such a situation, the null distribution of $2\log(T)$ would be approximately chi-square with one degree of freedom (see for instance Chapter 4 in Davison , 2008). However, these regularity assumptions are not satisfied, as the null value of the parameter $p$ lies at the boundary of the parameter space. We will thus also obtain critical values for the test via simulations.

**Note**: It should be noted that this test can be adapted to the case where an estimate of the error probability is given for each read. In this case, define $\varepsilon_{i,j}$ to be the estimate of the error probability for the $j$-th read from lane $i$. Let $Y_{i,j}$ be a 0-1 random variable, such that $Y_{i,j} = 0$ when the $j$-th read from lane $i$ indicates the major allele and $Y_{i,j} = 1$ when the $j$-th read from lane $i$ indicates the prospective minor allele. Define $\mathbf{Y_i} = (Y_{i,1}, Y_{i,2}, \ldots, Y_{i,R_i})$. The likelihood of the reads from lane $i$ under the null hypothesis is given by

$$L_0(\mathbf{y}_i|R_i = r_i) = \prod_{j=1}^{r_i}(1-\varepsilon_{i,j})^{1-y_{i,j}}\varepsilon_{i,j}^{y_{i,j}}.$$

Under the alternative hypothesis, the likelihood of the reads from lane $i$ conditional on the number of individuals in the lane with the minor allele is given by

$$L_1(\mathbf{y}_i|A_i = a_i, R_i = r_i) = \prod_{j=1}^{r_i}[1-q(a_i;\varepsilon_{i,j})]^{1-y_{i,j}}[q(a_i;\varepsilon_{i,j})]^{y_{i,j}}.$$

Arguing as in the case where the probability of error is constant, the likelihood ratio is given by

$$T = \frac{\max_p \prod_{i=1}^{k}\sum_{a_i=0}^{m}\left[\binom{m}{a_i}p^{a_i}(1-p)^{m-a_i}\left(\prod_{j=1}^{r_i}[1-q(a_i;\varepsilon_{i,j})]^{1-y_{i,j}}[q(a_i;\varepsilon_{i,j})]^{y_{i,j}}\right)\right]}{\prod_{i=1}^{k}\prod_{j=1}^{r_i}(1-\varepsilon_{i,j})^{1-y_{i,j}}\varepsilon_{i,j}^{y_{i,j}}}.$$

# 4 A Simple Test for the Presence of a Minor Allele - The Maximum Test

An intuitive alternative to the LR test is to accept that there is a minor allele when any of the lanes gives a sufficiently large number of reads of such an allele. Define $U_k = \max_{1\le i\le k} X_i$, i.e., $U_k$ is the maximum number of reads of a prospective minor allele from a lane. Under $H_0$ and the model presented in Section 2, $U_k$ is the maximum of $k$ independent observations from a Poisson$(\lambda\varepsilon)$ distribution. To control

the probability of a type I error, the critical value, $u_k$, is defined to be the smallest integer satisfying

$$P(U_k \leq u_k|H_0) \geq 1-\alpha \Rightarrow P(X_i \leq u_k|H_0)^k \geq 1-\alpha \Rightarrow P(X_i \leq u_k|H_0) \geq \sqrt[k]{1-\alpha}. \quad (6)$$

Hence, we can take $u_k$ to be the $\sqrt[k]{1-\alpha}$ quantile of the Poisson$(\lambda\varepsilon)$ distribution. We reject $H_0$ if and only if $U_k > u_k$. Since $\sqrt[k]{1-\alpha}$ is increasing in $k$, it follows that $u_k$ is non-decreasing in $k$ (given the remaining parameters are fixed). Also, the critical value is non-decreasing in the error rate.

Another way of obtaining this test is to construct a multiple hypothesis test that controls the familywise error using the $k$ test statistics $X_1, X_2, \ldots, X_k$. For small $\alpha$, the critical value can be accurately approximated using a significance level of $\alpha/k$, i.e., using the Bonferroni procedure. The resulting test controls for multiplicity across lanes, but not across loci. One might therefore use an additional procedure that controls the familywise error or the false discovery rate across loci (Benjamini and Hochberg , 1995).

In Section 5, we investigate this test under more general assumptions concerning the sequencing errors.

## 4.1 Estimation of the Power for Small $p$

Now we consider the distribution of the statistic $U_k$ used with the maximum test. Our focus is on the alternative hypothesis that the minor allele frequency is $p$, where $p$ is small. The number of individuals with the minor allele in the sample, $B = \sum_{i=1}^{k} A_i$, has a binomial distribution with parameters $n$ and $p$. This can be approximated by the Poisson$(np)$ distribution.

**Lemma 1**. *Suppose the pool size is fixed and ignore the possibility of errors. Assume there are b individuals with the minor allele in the sample. The distribution of the maximum number of reads of a minor allele across all lanes is stochastically smallest when one individual with the minor allele appears in each of b lanes.*

**Proof**. Assume there are $b$ individuals with the minor allele in the sample, who appear in $l$ lanes, where $l \leq b$. Let $\mathbf{1}^{(b)}$ be a vector of ones of length $b$. Let $\mathbf{a}^{(b)} = (a_1^{(b)}, a_2^{(b)}, \ldots, a_l^{(b)})$, where the $a_i^{(b)}$ are positive integers that sum to $b$. Suppose we relabel the lanes in order of the number of individuals with the minor allele (lane 1 has the maximum number of individuals with the minor allele). Similarly, the individuals with the minor allele are labeled according to the lane number, i.e., if $\mathbf{A} = \mathbf{a}^{(b)}$, then the individuals with the minor allele in lane 1 are numbered from 1 to $a_1^{(b)}$, those in lane 2 are numbered from $a_1^{(b)}+1$ to $a_1^{(b)}+a_2^{(b)}$ etc. Let $c_i^{(b)} = \sum_{j=1}^{i} a_j^{(b)}$ be the number of individuals with the minor allele in the first $i$ lanes labeled in this way. Set $c_0^{(b)} = 0$.

Let $V_j$ be the number of reads from the $j$-th individual with the minor allele. Hence, $V_j \sim \text{Poisson}(\lambda/m)$. We show that $U_l|\mathbf{A} = \mathbf{a}^{(b)}$ stochastically dominates $U_b|\mathbf{A} = \mathbf{1}^{(b)}$, i.e., $P(U_l \geq u|\mathbf{A} = \mathbf{a}^{(b)}) \geq P(U_b \geq u|\mathbf{A} = \mathbf{1}^{(b)})$. The number of reads of the minor allele from lane $i$, $1 \leq i \leq l$, when $\mathbf{A} = \mathbf{a}^{(b)}$ is $Y_i = \sum_{j=c_{i-1}^{(b)}+1}^{c_i^{(b)}} V_j$. As the $V_j$ are non-negative, for any set of realizations of $V_{c_{i-1}^{(b)}+1}, V_{c_{i-1}^{(b)}+2}, \ldots, V_{c_i^{(b)}}$, this sum will be greater than or equal to $\max_{c_{i-1}^{(b)}+1 \leq j \leq c_i^{(b)}}\{V_j\}$. Hence, $\sum_{j=c_{i-1}^{(b)}+1}^{c_i^{(b)}} V_j$ stochastically dominates $\max\{V_{c_{i-1}^{(b)}+1}, V_{c_{i-1}^{(b)}+2}, \ldots, V_{c_i^{(b)}}\}$. Taking the maximum of the $Y_i$, $\max_{1 \leq i \leq l}\{Y_i|\mathbf{A} = \mathbf{a}^{(b)}\}$ stochastically dominates $\max_{1 \leq j \leq b}\{V_j\}$, which is the maximum number of reads of the minor allele in a lane when all $b$ individuals with the minor allele appear in different lanes. □

We can thus obtain a lower bound on the power of the test by assuming that all the individuals with the minor allele appear in separate lanes. When the expected number of individuals with the minor allele is small compared to the number of lanes, the probability of having several individuals with this allele in any lane is small. Let $D$ denote the event that the alternative $H_A : p > 0$ is accepted. Conditioning on the number of individuals with the minor allele in the sample

$$P[D] = \sum_{b=0}^{\infty} P[D|B = b]P(B = b) \geq \sum_{b=1}^{\infty} P[D|B = b]P(B = b). \tag{7}$$

The approximation obtained by replacing the inequality by an equality is accurate, since the probability of accepting the alternative given that no individuals with the minor allele appear in the sample is simply the significance level of the test (which by assumption is small). For $b \geq 1$,

$$P[D|B = b] = P(U > u_c|B = b) = 1 - P(U_k \leq u_k|B = b) \geq 1 - P(V_1 \leq u_k)^b, \tag{8}$$

where $V_1$ is the number of reads from an individual with the minor allele. Replacing this inequality by an equality leads to an accurate approximation when $1/m$ is large compared with $\varepsilon$ (i.e., the probability that the maximum number of reads of the prospective allele comes from a lane where the minor allele is actually present is close to 1) and the probability of obtaining multiple copies of the minor allele in a lane is small. It follows that $P[D|B = b] \approx 1 - r_k^b$, where $r_k = \sum_{j=0}^{u_k} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}$.

Let $\mu$ be the expected number of individuals with the minor allele in the sample, i.e., $\mu = np = mkp$. It follows that for small $p$

$$P[D] \approx \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b[1 - r_k^b]}{b!} = \sum_{b=1}^{\infty} \frac{e^{-\mu}\mu^b}{b!} - \sum_{b=1}^{\infty} \frac{e^{-\mu}(\mu r_k)^b}{b!}$$

$$= (1-e^{-\mu}) - e^{-\mu(1-r_k)} \sum_{j=1}^{\infty} \frac{e^{-r_k\mu}(\mu r_k)^b}{b!}$$
$$= (1-e^{-\mu}) - e^{-\mu(1-r_k)}(1-e^{-r_k\mu}) = 1-e^{-\mu(1-r_k)}.$$

## 4.2 Optimization of the Pool Size for a Fixed Number of Lanes

The goal is to maximize the power of the maximum test at a fixed significance level.

**Result 1**. *Suppose that the number of lanes, k, is fixed. Let $\overline{m}_k$ be the unique solution of the following equation for m:*

$$0 = \sum_{j=u_k+1}^{\infty} \frac{e^{-\lambda/m}(\lambda/m)^j[1+\lambda/m-j]}{j!}. \tag{9}$$

*Denote the pool size that maximizes the approximation of the power function by $m_k^*$. When $\overline{m}_k$ is a positive integer, $m_k^* = \overline{m}_k$, $m_k^* = 1$ when $\overline{m}_k < 1$, otherwise $m_k^*$ is either the integer part of $\overline{m}_k$, $(\lfloor \overline{m}_k \rfloor)$ or one plus this integer part, $\lceil \overline{m}_k \rceil$, whichever minimizes the function*

$$f_k(m;p,\alpha) = e^{-\mu(1-r_k)} = \exp\left[-mkp \sum_{j=u_k+1}^{\infty} \frac{e^{-\lambda/m}(\lambda/m)^j}{j!}\right]. \tag{10}$$

**Proof**. Note that minimizing $f_k$ is equivalent to maximizing the approximation of the power function. Assuming that $m$ can take any positive real value and differentiating $f_k$ with respect to $m$,

$$f_k'(m;p,\alpha) = -f_k(m;p,\alpha)kp \sum_{j=u_k+1}^{\infty} \frac{e^{-\lambda/m}(\lambda/m)^j[1+\lambda/m-j]}{j!}. \tag{11}$$

Hence, $f_k$ has an extreme point if and only if Equation (9) is satisfied.

Note that $\lambda/\overline{m}_k > u_k$, since the first term of the sum in Equation (9) must be positive. Hence, at the optimal pool size (maximizing over the set of positive reals), the expected number of reads from an individual is greater than the critical value for the test.

Since $\eta = g(m) = \lambda/m$ is a monotone function of $m$ mapping $(0,\infty)$ onto $(0,\infty)$, to show that there is a unique solution of Equation (9), it suffices to show that there is a unique solution of the following equation in $\eta$:

$$h_k(\eta) = \sum_{j=u_k+1}^{\infty} P(W=j)[\eta+1-j] = 0, \tag{12}$$

where $W \sim \text{Poisson}(\eta)$. Differentiating $h_k$ and rearranging, we obtain $h_k'(\eta) = (\eta-u_k)P(W=u_k)$. If $\eta \le u_k$, then $h_k(\eta) < 0$. Also, $h_k'(\eta) > 0$ for all $\eta > u_k$ and

$\lim_{\eta\to\infty} h_k(\eta) > 0$. It follows from the continuity of $h_k$ that there is a unique value $\eta_k$ which satisfies Equation (12). Hence, there is a unique extreme point of the function $f_k(m)$ when $m = \overline{m}_k = \lambda/\eta_k$.

From the above analysis, when $m < \overline{m}_k$, i.e., $\eta > \eta_k$, $f_k'(m; p_0, \alpha) < 0$ and when $m > \overline{m}_k$, i.e., $\eta < \eta_k$, $f_k'(m; p_0, \alpha) > 0$. Hence, the extreme point of $f_k$ is a minimum as required.

If $\overline{m}_k$ is a positive integer, then from the form of $f_k$, it must be the integer pool size which maximizes the approximation of the power given above. If $\overline{m}_k < 1$, then a pool size of 1 maximizes the approximation of the power. If $\overline{m}_k$ is a non-integer greater than 1, then the integer pool size that maximizes the approximation of the power must be either $\lfloor \overline{m}_k \rfloor$ or $\lceil \overline{m}_k \rceil$. □

As argued above, this estimate of power is accurate when the frequency of the minor allele is small enough to ensure that it is unlikely that several individuals with the minor allele appear in the same pool. Moreover, simulations indicate that in practice this estimate of the power is accurate even when this assumption does not hold. Note that $\overline{m}_k$ is independent of $p$. Hence, it is reasonable to use $m_k^*$ as the optimal pool size for the detection of SNPs when the minor allele has a low frequency, i.e., as the asymptotically optimal pool size.

There may be two neighboring integers maximizing this estimate of the power. This occurs only when $f_k(\lfloor \overline{m}_k \rfloor; p, \alpha) = f_k(\lceil \overline{m}_k \rceil; p, \alpha)$. In this case, we take the optimal pool size to be $\lfloor \overline{m}_k \rfloor$. We now describe the relation between $m_k^*$ and the number of lanes. Recall that the critical value $u_k$ is non-decreasing in $k$.

**Result 2**. *a) The asymptotically optimal pool size for SNP detection depends on k only through the critical value $u_k$, b) $\overline{m}_k$ is non-increasing in k.*

**Proof**. To prove a), it suffices to show that when $u_k = u_{k+1} = c$, then $m_k^* = m_{k+1}^*$. In this case, from Equation (10) we obtain $f_{k+1}(m; p, \alpha) = [f_k(m; p, \alpha)]^{1+1/k}$. Hence, minimizing $f_k(m; p, \alpha)$ is equivalent to minimizing $f_{k+1}(m; p, \alpha)$.

To prove b), first note that when $u_{k+1} = u_k$, it follows from the above argument that $\overline{m}_{k+1} = \overline{m}_k$. Since $u_k$ is non-decreasing in $k$, it suffices to consider the case $u_{k+1} = u_k + 1$. From Equation (12),

$$\begin{aligned} h_k(\eta) - h_{k+1}(\eta) &= \sum_{j=u_k+1}^{\infty} P(W = j)[\eta + 1 - j] - \sum_{j=u_k+2}^{\infty} P(W = j)[\eta + 1 - j] \\ &= (\eta - u_k) P(W = u_k + 1), \end{aligned}$$

where $W \sim \text{Poisson}(\eta)$ and $\eta = \lambda/m$ is the expected number of reads from an individual in a pool of size $m$. By definition $h_k(\eta_k) = 0$ and, as argued previously, $\eta_k > u_k$. It follows that

$$h_{k+1}(\eta_k) = (u_k - \eta_k) P(W = u_k + 1) < 0. \tag{13}$$

Differentiating $h_{k+1}$, we obtain

$$h'_{k+1}(\eta) = [\eta - u_k - 1]P(W = u_k + 1). \tag{14}$$

Since the first term in the sum defining $h_{k+1}(\eta_{k+1})$ must be positive, there is no solution of $h_{k+1}(\eta) = 0$ with $\eta \in (0, u_k + 1)$. Let $\overline{\eta}_k = \max\{\eta_k, u_k + 1\}$. Thus $h_{k+1}(\overline{\eta}_k) < 0$. For $\eta > \overline{\eta}_k$, $h_{k+1}(\eta)$ is increasing in $\eta$ and $\lim_{\eta\to\infty} h_{k+1}(\eta) > 0$.

It follows that there is a unique positive solution, $\eta_{k+1}$, of the equation $h_{k+1}(\eta) = 0$ and $\eta_{k+1} > \eta_k$. Hence, $\overline{m}_{k+1} = \lambda/\eta_{k+1} < \overline{m}_k = \lambda/\eta_k$. □

The second result given above suggests the following conjecture:

**Conjecture 1**. *The asymptotically optimal pool size for detecting SNPs is non-increasing in the number of available lanes (all other parameters fixed).*

Since the critical value of the test is non-decreasing in the error rate, the proof of the second result above also suggests the following conjecture:

**Conjecture 2**. *The optimal pool size for detecting SNPs is non-increasing in the error rate (all other parameters fixed).*

These conjectures seem difficult to prove due to the complexity of dealing algebraically with integer parts when analyzing the functions $f_k$ and $f_{k+1}$. However, no counterexamples to these conjectures have been found in the numerous calculations that have been carried out.

Suppose a decrease in the error rate leads to a lower critical value. The optimal power to detect a rare minor allele must increase. This follows from observing that if we do not change the pool size, then the probability that the maximum number of reads from an individual with the minor allele exceeds the critical value must increase. A further increase in power is possible, if the change in the critical value changes the optimal pool size. Example 1 illustrates how an optimal pool size can be calculated in practice.

It should be noted that it is difficult to adapt this test to the case where the error probability is variable. In this case, $\varepsilon$ should be interpreted as an upper bound on the probability of an error (see also the results from the simulations presented in Section 5).

## 4.3 Other Problems

Having determined the asymptotically optimal pool size given a fixed number of lanes, we now consider the following two problems:

**1** Given a fixed number of lanes, what is the minimum minor allele frequency that can be detected with the required power?

**2** What is the minimum number of lanes that is required to detect a minor allele of given frequency with the required power?

First, assume that the number of available lanes, $k$, is fixed. Denote the minimum minor allele frequency that can be detected with the required power $\beta$ using a significance level of $\alpha$ by $p_{min}(k;\beta,\alpha)$. In order to do this, we need to solve $1-f_k(m_k^*;p,\alpha)=\beta$ as an equation in $p$. This leads to

$$\ln(1-\beta) = -m_k^* k p_{min}(k;\beta,\alpha)[1-r_k] \Rightarrow p_{min}(k;\beta,\alpha) = \frac{-\ln(1-\beta)}{m_k^* k[1-r_k]}. \qquad (15)$$

Now we allow the number of available lanes to vary. Denote the minimum number of lanes required to detect a minor allele of frequency $p$ with power $\beta$ by $k(p;\beta,\alpha)$. By definition

$$k(p;\beta,\alpha) = \min_k p_{min}(k;\beta,\alpha) \leq p. \qquad (16)$$

The problem is solvable, since $p_{min}(k;\beta,\alpha)$ can be found for $k=1,2,3,\ldots$ in turn. This calculation can be made more efficient using the fact that $m_k^*$ and $r_k$ only depend on $k$ through the critical value of the test, which often does not change over a wide range of $k$. This calculation is illustrated in Example 2.

**Example 1**. Suppose that 40 lanes are available, the mean number of reads per lane is 20 and the error rate is 0.01. We derive an optimal pool size for detecting SNPs using a significance level of 0.1% .

Under $H_0$, the number of reads of the prospective minor allele in a lane has a Poisson(0.2) distribution. The critical value for this test, $u_{40}$, is thus the $\sqrt[40]{0.999}$ quartile of this distribution. Hence, $u_{40}=4$. We thus conclude that the site is a SNP when there are at least 5 reads of a prospective minor allele in a lane.

To find the optimal pool size, we minimize the function

$$\begin{aligned} f_k(m;p,\alpha) &= e^{-\mu(1-r_k)} = \exp\left[-mkp\sum_{j=u_k+1}^{\infty}\frac{e^{-\lambda/m}(\lambda/m)^j}{j!}\right] \\ &= \exp\left[-40mp\sum_{j=5}^{\infty}\frac{e^{-20/m}(20/m)^j}{j!}\right]. \end{aligned}$$

This is equivalent to maximising

$$d(m) = m\left[1-\sum_{j=0}^{4}\frac{e^{-20/m}(20/m)^j}{j!}\right]. \qquad (17)$$

In practice, it is simplest to maximize this function by calculating $d(m)$ for integer values of $m$ using the fact that the expected number of reads, $20/m$, from an individual at the optimal pool size has to be greater than the critical value. It follows

that the optimal pool size cannot be greater than 5. We can find the optimal pool size by evaluating $d(m)$ at decreasing integer values starting at this upper bound. The optimal pool size is $m^*_{40} = 3$. The minimum frequency that is detectable with a power of 0.95 is given by

$$p_{min}(40; 0.95, 0.001) = \frac{-\ln(1-\beta)}{40 m^*_{40}[1 - r_{40}]}, \tag{18}$$

where $m^*_k = 3$, $u_{40} = 4$ and thus

$$q_{40} = e^{-20/3} \sum_{j=0}^{4} \frac{(20/3)^j}{j!} \approx 0.2056. \tag{19}$$

Hence,

$$p_{min}(40; 0.95, 0.001) = \frac{-\ln(0.05)}{120 \times 0.7944} \approx 0.0314. \tag{20}$$

Table 1 gives results for various numbers of lanes, $k$, and error rates, $\varepsilon$.

Table 1: Critical values, optimal pool sizes and minimum detectable frequencies with power 0.95 at a significance level of 0.1% . The mean read rate is 20.

| | $k = 16$ | $k = 40$ | $k = 80$ | $k = 120$ |
|---|---|---|---|---|
| $\varepsilon = 0.01$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.0637$ | $u_k = 4, m^* = 3,$ $p_{min} = 0.0314$ | $u_k = 4, m^* = 3,$ $p_{min} = 0.0157$ | $u_k = 4, m^* = 3,$ $p_{min} = 0.0105$ |
| $\varepsilon = 0.005$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.0637$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.0255$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.0127$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.00849$ |
| $\varepsilon = 0.002$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.0482$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.0193,$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.00964$ | $u_k = 3, m^* = 4,$ $p_{min} = 0.00849$ |
| $\varepsilon = 0.001$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.0482$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.0193$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.00964,$ | $u_k = 2, m^* = 6,$ $p_{min} = 0.00643$ |

**Example 2**. Suppose that a lane gives on average 20 reads with an error rate of 0.01. We wish to find the number of lanes that are required to discover a minor allele of frequency 1% with a power of 0.95 using a significance level of 0.1%. Suppose that up to 1000 lanes could be used.

The critical value for the test with $k$ lanes, $u_k$, is given by the $\sqrt[k]{0.999}$ quartile of the Poisson(0.2) distribution. It follows that for $k \leq 17$, $u_k = 3$, for $18 \leq k \leq 443$, $u_k = 4$, for $444 \leq k \leq 1000$, $u_k = 5$. By calculating $p_{min}(k; \beta, \alpha)$ for the largest value

of $k$ for which a particular critical value holds, we can first find the interval in which the solution must lie. For $k \leq 17$, $u_k = 3$ and $m_k^* = 4$. Hence,

$$p_{min}(17; 0.95, 0.001) = \frac{-\ln(0.05)}{68[1 - e^{-5}\sum_{j=0}^{3}\frac{5^j}{j!}]} \approx 0.0599. \tag{21}$$

Thus the number of lanes must be greater than 17. For $18 \leq k \leq 443$, $u_k = 4$ and $m_k^* = 3$. Hence,

$$p_{min}(443; 0.95, 0.001) = \frac{-\ln(0.05)}{443 \times 3[1 - e^{-20/3}\sum_{j=0}^{4}\frac{(20/3)^j}{j!}]} \approx 0.00284. \tag{22}$$

It follows that the minimum number of lanes, $k(0.01; 0.95, 0.001)$, must be between 18 and 443. The solution can be found by solving

$$p_{min}(k; 0.95, 0.001) = \frac{-\ln(0.05)}{3k[1 - e^{-20/3}\sum_{j=0}^{4}\frac{(20/3)^j}{j!}]} = 0.01. \tag{23}$$

This leads to

$$k = \frac{-\ln(0.05)}{3 \times 0.7944 \times 0.01} \Rightarrow k = 125.71. \tag{24}$$

Thus 126 lanes are required. With an Illumina genome sequencer an experiment involves 8 lanes, making it natural to choose the number of lanes to be a multiple of 8. Hence, we require 128 lanes, i.e., 16 gene sequencing experiments.

As expected, when the number of lanes increases, the minimum minor allele frequency $p_{min}(k; \beta, \alpha)$ that is detectable with a given power tends to decrease. From Equation (15), if the critical value of the test is constant over a range of $k$, then $p_{min}(k; \beta, \alpha)$ is inversely proportional to the number of lanes. Note that the test statistic comes from a discrete distribution and the significance level is controlled to be $\leq \alpha$. Due to this, whenever the critical value increases, the actual significance level of the test falls from close to $\alpha$ to something maybe relatively much smaller. This causes a fall in power. One could avoid this problem by rejecting the null hypothesis when $U = u_k$ with the appropriate probability to make the actual significance level equal to $\alpha$. However, due to philosophical issues related to such a procedure, we do not investigate such tests. Adopting a multiple testing approach based on the number of reads of the minor allele from each lane and the Benjamini-Hochberg procedure (see Benjamini and Hochberg , 1995) would give an increase in power, but does not solve the problem regarding the discrete nature of the test statistic.

# 5 Results from Simulations

To estimate the power of the test for a given read rate $\lambda = 20$, frequencies of the minor allele $p \in \{0.005, 0.01, 0.02, 0.05\}$, number of lanes $k \in \{16, 40, 80, 120\}$ and error probabilities $\varepsilon \in \{0.001, 0.002, 0.005, 0.01\}$, we simulated 10,000 tests for each case. In general, the number of individuals in a pool was varied from one to ten. It should be noted that the asymptotically optimal pool size in the problems considered is between three and six. The optimal pool size was taken to be the minimum pool size for which the maximum power was obtained.

We also investigated a scenario which could be thought of as illustrating the effect of amplifying a particular section of the genome when a small number of lanes (only 8 in this case) are available. In this case, it was assumed that the read rate is $\lambda = 60$.

It should be noted that the *LR* test requires numerical maximization to calculate the likelihood ratio statistic. The numerator of this expression (see Equation 5) was calculated at a grid of 20 equally spaced points $p = 0.003, 0.006, \ldots, 0.06$. Further simulations indicate that increasing the density of grid points did not not lead to any visible gain in power. We carried out simulations under three models:

**Model 1** Any errors in reading the major allele give the minor allele and vice versa (i.e., the model presented in Section 2).

**Model 2** Any errors in reading an allele always give the same base, which is neither the major nor the minor allele.

**Model 3** An error is equally likely to give any of the three other possible alleles.

Note that the maximum test assumes that the expected number of reads from a lane is known. For the purposes of the simulations, the maximum test was also adapted so that $\lambda$ is estimated from the data and then the appropriate critical value calculated from this estimate. The LR test does not require estimation of $\lambda$.

## 5.1 Adaptation of the Tests to More General Models

Assume now that reads of all four nucleotides can be obtained, the probability of an error and the minor allele frequency are small ($\leq 0.01$ and $\leq 0.1$, respectively) and the sample size is relatively large. Under these assumptions, we may assume that the major allele obtains the greatest total number of reads with probability essentially equal to one. Hence, we must decide which of the other three nucleotides is the putative minor allele.

Using the maximum test, the non-major allele giving the largest number of reads in a single lane is deemed to be the putative minor allele. For the LR test,

we consider the 3 non-major alleles in turn. Each time, we classify the reads into two groups: (a) reads of the allele under consideration, (b) all other reads (assumed to be reads of the major allele). We then calculate the realization of the LR test statistic corresponding to each of the non-major alleles. The putative minor allele is defined to be the one for which the maximum is achieved. If this statistic exceeds the appropriate critical value for the test, then we accept that the putative minor allele is present.

**Note 1**: Due to the fact that we are taking the maximum of three test statistics, there is an element of multiple testing in this approach. However, the three test statistics obtained in this way will be highly correlated, making the problem less serious. Furthermore, the proposed approach can be taken into account when simulating critical values.

**Note 2**: Under such a procedure, in addition to the standard type I and II testing errors, when a minor allele is present we may conclude that some other non-major allele is present. Such an error will be referred to as a type III error. The following condition must be satisfied for a type III error to be committed: the realization of the test statistic corresponding to the allele in question must be greater than both a) the critical value and b) the realization of the test statistic corresponding to the actual minor allele. For the maximum test, under Model 1 the probability that just the first condition is fulfilled is bounded above by $\alpha$. In general, we expect that the probability of a type III error will be small relative to the nominal significance level.

## 5.2 Power of the Tests under Model 1

The results presented in Tables 2 and 3 compare some of the theoretical and empirical results for the maximum test, together with the LR test, under the assumptions of Model 1. It should be noted that the power of the *LR* test is given for the empirically determined optimal pool size for that test (obtained by simulation with a maximum pool size of 10). The empirical power of the maximum test is very much in agreement with the theoretical calculations made in the previous section. The estimates from these simulations tend to be slightly greater than the theoretical estimates. This results from the facts that, firstly, these theoretical estimates were obtained by ignoring errors of the major allele being read as the minor allele

Table 2: Estimates of the maximum power and optimal pool sizes (in brackets) under Model 1 with mean read rate $\lambda = 20$, together with theoretical estimates for the maximum test (taken over pool sizes from 1 to 10, based on 10 000 simulations).

| **Theoretical** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $p = 0.01, \varepsilon = 0.01$ | 0.3752 (4) | 0.6145 (3) | 0.8514 (3) |
| $p = 0.01, \varepsilon = 0.005$ | 0.3752 (4) | 0.6915 (4) | 0.9048 (4) |
| $p = 0.01, \varepsilon = 0.002$ | 0.4628 (6) | 0.7885 (6) | 0.9553 (6) |
| $p = 0.02, \varepsilon = 0.01$ | 0.6097 (4) | 0.8514 (3) | 0.9779 (3) |
| $p = 0.02, \varepsilon = 0.005$ | 0.6097 (4) | 0.9048 (4) | 0.9909 (4) |
| $p = 0.02, \varepsilon = 0.002$ | 0.7114 (6) | 0.9553 (6) | 0.9980 (6) |
| $p = 0.05, \varepsilon = 0.01$ | 0.9048 (4) | 0.9915 (3) | 0.9999 (3) |
| $p = 0.05, \varepsilon = 0.005$ | 0.9048 (4) | 0.9972 (4) | 1.0000 (4) |
| $p = 0.05, \varepsilon = 0.002$ | 0.9553 (6) | 0.9996 (6) | 1.0000 (6) |
| **$\lambda$ known** | $k = 16$ | $k = 40$ | $k = 80$ |
| $p = 0.01, \varepsilon = 0.01$ | 0.3864 (4) | 0.6252 (3) | 0.8489 (3) |
| $p = 0.01, \varepsilon = 0.005$ | 0.3825 (4) | 0.6973 (4) | 0.9065 (4) |
| $p = 0.01, \varepsilon = 0.002$ | 0.4733 (6) | 0.7995 (7) | 0.9583 (7) |
| $p = 0.02, \varepsilon = 0.01$ | 0.6237 (5) | 0.8624 (3) | 0.9774 (3) |
| $p = 0.02, \varepsilon = 0.005$ | 0.6218 (4) | 0.9133 (4) | 0.9923 (5) |
| $p = 0.02, \varepsilon = 0.002$ | 0.7195 (6) | 0.9603 (6) | 0.9984 (7) |
| $p = 0.05, \varepsilon = 0.01$ | 0.9190 (4) | 0.9934 (3) | 1.0000 (3) |
| $p = 0.05, \varepsilon = 0.005$ | 0.9074 (4) | 0.9976 (4) | 1.0000 (2) |
| $p = 0.05, \varepsilon = 0.002$ | 0.9653 (6) | 0.9997 (8) | 1.0000 (3) |
| **$\lambda$ unknown** | $k = 16$ | $k = 40$ | $k = 80$ |
| $p = 0.01, \varepsilon = 0.01$ | 0.3525 (5) | 0.6212 (3) | 0.8531 (3) |
| $p = 0.01, \varepsilon = 0.005$ | 0.3784 (4) | 0.6921 (4) | 0.9098 (4) |
| $p = 0.01, \varepsilon = 0.002$ | 0.4688 (7) | 0.7981 (7) | 0.9598 (6) |
| $p = 0.02, \varepsilon = 0.01$ | 0.5937 (4) | 0.8586 (3) | 0.9762 (3) |
| $p = 0.02, \varepsilon = 0.005$ | 0.6108 (4) | 0.9061 (4) | 0.9911 (4) |
| $p = 0.02, \varepsilon = 0.002$ | 0.7207 (6) | 0.9567 (7) | 0.9978 (4) |
| $p = 0.05, \varepsilon = 0.01$ | 0.9004 (4) | 0.9926 (4) | 1.0000 (3) |
| $p = 0.05, \varepsilon = 0.005$ | 0.9090 (4) | 0.9977 (4) | 1.0000 (3) |
| $p = 0.05, \varepsilon = 0.002$ | 0.9627 (6) | 0.9998 (6) | 1.0000 (3) |

and, secondly, the estimates were derived to be tight lower bounds on the power when the minor allele frequency is small.

Table 3: Estimates of the maximum power and optimal pool sizes (in brackets) under Model 1 for the LR test with mean read rate $\lambda = 20$ (taken over pool sizes from 1 to 10, based on 10 000 simulations).

| **LR Test** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $p = 0.01, \varepsilon = 0.01$ | 0.3450 (4) | 0.6398 (4) | 0.8684 (5) |
| $p = 0.01, \varepsilon = 0.005$ | 0.3891 (5) | 0.7097 (4) | 0.9127 (4) |
| $p = 0.01, \varepsilon = 0.002$ | 0.4833 (8) | 0.7811 (7) | 0.9576 (7) |
| $p = 0.02, \varepsilon = 0.01$ | 0.5738 (4) | 0.8846 (5) | 0.9889 (5) |
| $p = 0.02, \varepsilon = 0.005$ | 0.6416 (8) | 0.9328 (7) | 0.9978 (8) |
| $p = 0.02, \varepsilon = 0.002$ | 0.7446 (9) | 0.9687 (10) | 0.9993 (10) |
| $p = 0.05, \varepsilon = 0.01$ | 0.9253 (9) | 0.9997 (10) | 1.0000 (3) |
| $p = 0.05, \varepsilon = 0.005$ | 0.9607 (10) | 0.9999 (7) | 1.0000 (4) |
| $p = 0.05, \varepsilon = 0.002$ | 0.9805 (10) | 1.0000 (9) | 1.0000 (3) |

In the majority of cases, the theoretically derived optimal pool size gave the maximum power. The cases where the theoretically derived optimal pool size differed from the empirically determined optimal pool size can be split into two cases: i) a neighboring pool size gave a very similar power, ii) for a large minor allele frequency and number of lanes, the power of the test is essentially 1 for a large range of pool sizes. In this case, the optimal pool size from the simulation was defined to be the smallest pool size for which the maximum power is achieved and thus may be much smaller than the theoretically determined optimal pool size. In these cases, the asymptotically optimal pool size also gave an estimated power of 1.

It can be seen that when the expected number of reads must be estimated and the number of lanes is low, the power of the maximum test is generally lower. This is particularly noticeable when the error rate is relatively high. However, when the number of lanes available is large, the need to estimate the read rate has no visible effect on the power, since an accurate estimate of $\lambda$ will be obtained. However, the choice of the appropriate pool size may be a problem in this case.

Figures 1 and 2 give a comparison of the powers of the maximum test and the *LR* test. In general, the powers of the two tests are comparable. However, a few important details may be observed.

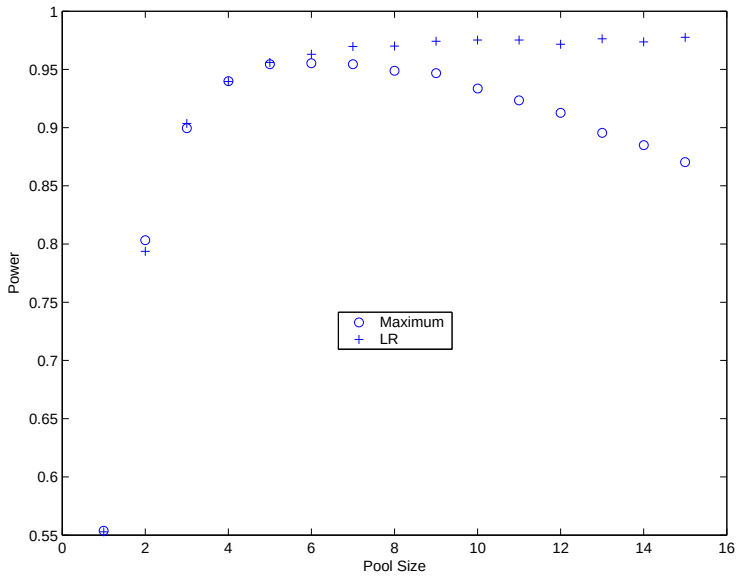


Figure 1: Comparison of the power of the tests for various pool sizes, $p = 0.01$, $k = 80$, $\varepsilon = 0.001$, $\alpha = 0.001$, $\lambda = 20$. Based on 10,000 simulations.

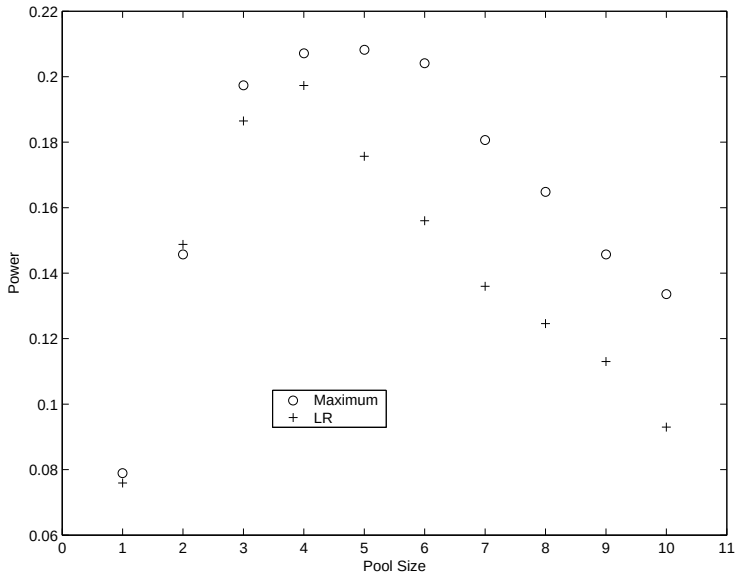


Figure 2: Comparison of the power of the maximum test and the *LR* test for various pool sizes, $p = 0.005$, $k = 16$, $\varepsilon = 0.01$, $\alpha = 0.001$, $\lambda = 20$. Based on 10,000 simulations.

**1** Using the asymptotically optimal pool size, when the expected number of individuals with the minor allele is small (i.e., when $p$ and $k$ are relatively small), the maximum test works well compared to the *LR* test. In such cases, the maximum number of reads of the putative minor allele contains a very large percentage of the information regarding the presence of a minor allele (see Figure 2).

**2** The discrete nature of the test statistic for the maximum test means that the power is not a smooth function of the parameters. When an increase in the number of lanes available, $k$, leads to an increase in the critical value of the test, the power of the maximum test may fall. For example, when $p = 0.005$, $k = 80$ and $\varepsilon = 0.002$, the power obtained using the maximum test is greater than the power obtained using the *LR* test. However, when the number of lanes is increased to $k = 120$ (which increases the critical value for the maximum test), the power obtained by the *LR* test is noticeably greater.

**3** When $km_k^*p$ is relatively large (i.e., more than a few individuals with the minor allele are expected at the optimal pool size), the empirical optimal pool size for the *LR* test is greater than the asymptotically optimal pool size. In such cases, for pool sizes up to the asymptotically optimal pool size, the two tests have very similar powers. For larger pool sizes, the power of the maximum test slowly decreases, whilst the power of the *LR* test seems to increase marginally before plateauing.

Of course, in practice, the pool size cannot be chosen to depend on the (unknown) frequency of the minor allele.

## 5.3 Significance Level of the Tests under Model 1

Table 4 gives the actual significance level of the maximum test when the nominal significance level is 5% and estimates are based on averaging over a hundred thousand simulations (10,000 for each pool size between 1 and 10). It should be noted that from the form of the maximum test, the actual significance level is independent of the pool size. The actual significance level $\overline{\alpha}$ can be calculated using $\overline{\alpha} = 1 - P(X \leq u_k)^k$, where $X \sim$Poisson$(\lambda\varepsilon)$, $u_k$ is the critical value for the test and $k$ is the number of lanes. It should be noted that when an increase in the number of lanes results in an increase in the critical value of the test, the actual significance level may fall by a large factor. This is noticeable when the number of lanes increases from 40 to 80 and the probability of an error is either 0.01 or 0.002.

The need to estimate the read rate, $\lambda$, generally has little influence on the actual significance level of the maximum test. The case $k = 40, \varepsilon = 0.01$ may indicate

Table 4: Empirical significance levels with mean read rate $\lambda = 20$ and nominal significance level 5%, together with the theoretical significance level for the maximum test. Since the actual significance level is independent of the pool size for the maximum test, the estimates are obtained by averaging over 100 000 simulations. The values given for the LR test are the maximum empirical significance levels observed. It should be noted that for the LR test there is a clear positive correlation between the pool size and the empirical significance level.

| **Maximum Test - Theoretical** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $\varepsilon = 0.01$ | 0.0182 | 0.0449 | 0.0045 |
| $\varepsilon = 0.005$ | 0.0025 | 0.0062 | 0.0123 |
| $\varepsilon = 0.002$ | 0.0124 | 0.0307 | 0.0008 |
| **Maximum Test - Empirical** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.0185 | 0.0445 | 0.0045 |
| $\varepsilon = 0.005$ | 0.0026 | 0.0063 | 0.0120 |
| $\varepsilon = 0.002$ | 0.0120 | 0.0310 | 0.0009 |
| **Maximum Test - $\lambda$ unknown** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.0181 | 0.0380 | 0.0049 |
| $\varepsilon = 0.005$ | 0.0024 | 0.0066 | 0.0121 |
| $\varepsilon = 0.002$ | 0.0132 | 0.0301 | 0.0008 |
| **LR Test - Empirical** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.0111 | 0.0125 | 0.0121 |
| $\varepsilon = 0.005$ | 0.0056 | 0.0106 | 0.0092 |
| $\varepsilon = 0.002$ | 0.0097 | 0.0085 | 0.0040 |

that when the actual significance level is close to the nominal significance level, then estimation of $\lambda$ may somewhat reduce the actual significance level. When the actual significance level is low compared to the nominal level, the conditional probability of the number of errors in a lane exceeding the critical value will be positively associated with the mean number of reads observed in the lanes. This is due to the fact that the critical value will not, in general, be affected by the mean number of reads. When the actual significance level is close to the nominal significance level, an increase in the mean number of reads may lead to an increase in the critical value, which results in a lower conditional probability of rejecting the null hypothesis.

It should be noted that both tests are clearly conservative. When Model 1 is valid the maximum test is conservative by definition, due to the discrete nature of

the test statistic. The actual significance level for the LR test is positively correlated with the pool size, but is much less affected by changes in the number of lanes and the error rate than the maximum test is.

## 5.4 The Case of a Small Number of Lanes and a High Read Rate

This scenario may well reflect practical cases where a small number of lanes are available and a particular section of the genome is of interest. In this case, PCR amplification may be applied to that section of the genome, resulting in a particularly high read rate.

Table 5 and Figure 3 compare the theoretical and empirical power of the maximum test, together with the LR test. It should be noted that when the minor allele frequency is 5%, it is quite probable that at the asymptotically optimal pool size there will be more than one individual with the minor allele in at least one of the lanes. However, the theoretical estimate of the power of the maximum test is still reasonably accurate. It should be noted that in such cases the power of the maximum test is close to one. Essentially, the estimate of the power of the maximum test is based on calculating an upper bound on the logarithm of the probability of not detecting a minor allele. Hence, when this logarithm is large and negative, its estimate does not have to be accurate in order to give an accurate estimate of the power.

In all the cases the power of the LR test is slightly greater than the power of the maximum test. Figure 3 illustrates the power of the tests for pool sizes up to 20. It can be seen that in this case the power of both tests seem to plateau once the asymptotically optimal pool size has been exceeded.

## 5.5 Effects of Deviations from the Simple Model

A comparison of the powers of the tests under each of the models stated at the beginning of Section 5 is presented in Table 6. In general, the powers obtained under Model 2 are minimally lower than under Model 1. In the case of the maximum test, they are very similar to the theoretical estimates of power. This slight fall in power is due to the fact that mistakes in reading the major allele do not result in a read of the minor allele.

Under Model 2, we are interested in the probability of stating that an allele which is not the minor allele is actually present (a type III error). These probabilities were all below 0.001 (the nominal significance level) and tend to decrease as: a) the number of lanes increases, b) the minor allele frequency increases and c) the probability of an error decreases.

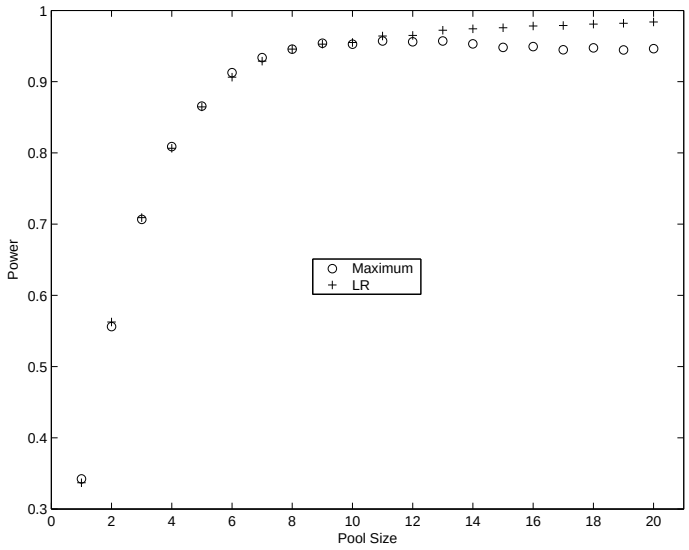


Figure 3: Comparison of the power of the maximum test and the *LR* test for various pool sizes, $p = 0.05$, $k = 8$, $\varepsilon = 0.005$, $\alpha = 0.001$, $\lambda = 60$. Based on 10,000 simulations.

Table 5: Estimates of the maximum power and optimal pool sizes under Model 1 with mean read rate $\lambda = 60$ and 8 lanes, together with theoretical estimates for the maximum test (taken over pool sizes from 1 to 20, based on 10 000 simulations).

| **Theoretical** | $\varepsilon = 0.01$ | $\varepsilon = 0.005$ | $\varepsilon = 0.002$ |
|---|---|---|---|
| $p = 0.02$ | 0.6213 (8) | 0.6814 (10) | 0.7561 (12) |
| $p = 0.05$ | 0.9117 (8) | 0.9427 (10) | 0.9706 (12) |
| **Maximum Test** | $\varepsilon = 0.01$ | $\varepsilon = 0.005$ | $\varepsilon = 0.002$ |
| $p = 0.02$ | 0.6484 (8) | 0.7029 (10) | 0.7707 (13) |
| $p = 0.05$ | 0.9318 (9) | 0.9579 (11) | 0.9813 (14) |
| **LR Test** | $\varepsilon = 0.01$ | $\varepsilon = 0.005$ | $\varepsilon = 0.002$ |
| $p = 0.02$ | 0.6386 (9) | 0.7257 (15) | 0.8099 (20) |
| $p = 0.05$ | 0.9649 (20) | 0.9850 (20) | 0.9931 (20) |

Note that when there is no minor allele, Model 2 is identical to Model 1. Hence, the actual significance levels of the maximum and the *LR* test will be the same under Model 2 as under Model 1.

When there is no minor allele, the maximum of the number of reads from the non-major alleles taken over all the lanes is stochastically dominated by the

Table 6: Comparison of the results obtained under the three models with mean read rate $\lambda = 20$ (taken over pool sizes from 1 to 10, based on 10 000 simulations). Unless otherwise stated, the figures given are powers and optimal pool sizes (given in brackets) when $p = 0.02$. The significance levels are given for a nominal significance level of 5%. Note that the actual significance levels are identical for Models 1 and 2.

| **Theoretical** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $\varepsilon = 0.01$ | 0.6097 (4) | 0.8514 (3) | 0.9779 (3) |
| $\varepsilon = 0.002$ | 0.7114 (6) | 0.9553 (6) | 0.9980 (6) |
| Sig. level, $\varepsilon = 0.01$ | 0.0182 | 0.0449 | 0.0045 |
| **Maximum Test - Model 1** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.6237 (5) | 0.8624 (3) | 0.9774 (3) |
| $\varepsilon = 0.002$ | 0.7195 (6) | 0.9603 (6) | 0.9984 (7) |
| Sig. level, $\varepsilon = 0.01$ | 0.0185 | 0.0445 | 0.0045 |
| **Maximum Test - Model 2** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.6017 (4) | 0.8481 (3) | 0.9784 (3) |
| $\varepsilon = 0.002$ | 0.7116 (6) | 0.9571 (7) | 0.9985 (6) |
| **Maximum Test - Model 3** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.6188 (4) | 0.8555 (3) | 0.9778 (3) |
| $\varepsilon = 0.002$ | 0.7205 (6) | 0.9599 (6) | 0.9986 (7) |
| Sig. level, $\varepsilon = 0.01$ | 0.0022 | 0.0054 | 0.0002 |
| **LR Test - Model 1** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.5738 (4) | 0.8846 (5) | 0.9889 (5) |
| $\varepsilon = 0.002$ | 0.7446 (9) | 0.9687 (10) | 0.9993 (10) |
| Sig. level, $\varepsilon = 0.01$ | 0.0111 | 0.0125 | 0.0121 |
| **LR Test - Model 2** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.5592 (3) | 0.8647 (4) | 0.9839 (4) |
| $\varepsilon = 0.005$ | 0.7285 (8) | 0.9623 (10) | 0.9990 (9) |
| **LR Test - Model 3** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.6792 (7) | 0.9513 (9) | 0.9981 (9) |
| $\varepsilon = 0.005$ | 0.7935 (10) | 0.9835 (10) | 0.9999 (9) |
| Sig. level, $\varepsilon = 0.01$ | 0.0263 | 0.0174 | 0.0239 |

corresponding maximum under Models 1 and 2. Hence, the actual significance level for the maximum test is lower under Model 3 than under Model 1 or 2. It

should be noted that the empirically derived actual significance levels of the *LR* test under Model 3 are similar to the ones obtained under Models 1 and 2 (and even sometimes somewhat higher). This may be due to the fact that the likelihoods of the data under both hypotheses will tend to be lower under Model 3 than under Model 1 and hence the distribution of the test statistic may well be relatively unaffected.

The estimated power of the maximum test under Model 3 is similar to (marginally lower than) the power obtained under Model 1. The probability of a type III error is lower than under Model 2, as the number of reads of each of the non-present alleles is stochastically dominated by the total number of errors obtained under Model 2, which are assumed to give the same base.

The powers obtained under Model 3 using the *LR* test are visibly higher than under Models 1 and 2. This increase in power seems to be of the same order as the increase obtained under the maximum test when the probability of an error is decreased by a factor of three. It should be noted that, in practice, sequencing errors may not be independent and may well point to the same incorrect nucleotide.

## 5.6 Effect of Inaccurate Estimation of the Error Rate

Table 7 illustrates the effect of overestimating the error rate as 0.01 on the power and actual significance level of the maximum test and the LR test. Comparing the results with those given in Tables 2 and 3, the power is comparable to the case where the true error rate is 0.01 and estimated correctly. In addition, the test becomes very conservative. Hence, it can be seen that the maximum test can be adapted to a variable error rate by using an upper bound on the probability of error. Although some power may be lost, the type I error rate is controlled. Similar results were obtained for the LR test. However, as mentioned before, the LR test can be easily adapted to the case when quality scores are given for each read.

Table 8 illustrates the effect of underestimating the error rate on the power and actual significance level of the maximum test and the LR test. In this case, there is no control of the type I error rate. For both tests, the actual significance level is increasing in the number of lanes used. In the case of the LR test, the actual significance level is also increasing in the pool size.

## 5.7 Effect of the Overdispersion of the Number of Reads

As might be expected, an increase in the variability of the number of reads from a lane has a negative effect on the power of the tests (see Table 9). This is more visible in the case of the maximum test, which explicitly assumes that the number of reads from a lane comes from the Poisson distribution. When the variance of the number

Table 7: Effect of overestimation of the error probabilities, $\lambda = 20$ (pool sizes from 1 to 10, based on 10 000 simulations). Unless otherwise stated, the figures given are powers and optimal pool sizes (given in brackets) for the given minor allele frequency. The significance levels are given for a nominal significance level of 5% and 0.1% . The true error rate is 0.001, the assumed error rate is 0.01.

| **Maximum Test** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $p = 0.01$ | 0.3758 (4) | 0.6193 (3) | 0.8508 (3) |
| $p = 0.02$ | 0.6094 (4) | 0.8541 (3) | 0.9787 (3) |
| $p = 0.05$ | 0.9124 (4) | 0.9925 (3) | 1.0000 (3) |
| Sig. level, $\alpha = 0.05$ | 0.00003 | 0.00006 | 0.00001 |
| Sig. level, $\alpha = 0.001$ | 0.00000 | 0.00000 | 0.00000 |
| **LR Test** | $k = 16$ | $k = 40$ | $k = 80$ |
| $p = 0.01$ | 0.3388 (4) | 0.6382 (4) | 0.8564 (4) |
| $p = 0.02$ | 0.5677 (4) | 0.8716 (4) | 0.9866 (4) |
| $p = 0.05$ | 0.8950 (4) | 0.9979 (6) | 1.0000 (4) |
| Sig. level, $\alpha = 0.05$ | 0.00000 | 0.00000 | 0.00000 |
| Sig. level, $\alpha = 0.001$ | 0.00000 | 0.00000 | 0.00000 |

of reads increases, the probability that the number of reads from an individual with the minor allele exceeds the critical value falls (since at the optimal pool size the critical value for the test must be smaller than the expected number of reads from an individual). This results in a fall in power. On the other hand, the probability that the number of errors from a lane exceeds the critical value increases as the variance of the number of reads increases. Hence, the actual significance level of the test is greater than the actual significance level of the test under Model 1. Note that in the case where $k = 40$ and $\varepsilon = 0.01$, under Model 1 the actual significance level is close to the nominal significance level. When the variance in the number of reads is increased, the actual significance level may exceed the nominal significance level. The LR test is less affected by the distribution of the number of reads, since the LR statistic is calculated conditional on the number of reads from each lane. Thus there is no explicit assumption regarding the distribution of the number of reads from each lane. When the variance of the number of reads increases, the probability of a small number of reads from a lane containing an individual with the minor allele increases, which will decrease the power of the test. However, since the LR test does not employ a fixed threshold rule, it is more flexible with regard to variation in

Table 8: Effect of underestimation of the error probabilities, $\lambda = 20$ (pool sizes from 1 to 10, based on 10 000 simulations). Unless otherwise stated, the figures given are powers and optimal pool sizes (given in brackets) for the given minor allele frequency. The significance levels are given for a nominal significance level of 5% and 0.1% (maximum significance levels are given for the LR test) . The true error rate is 0.01, the assumed error rate is 0.001.

| **Maximum Test** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $p = 0.01$ | 0.4890 (6) | 0.8115 (6) | 0.9639 (6) |
| $p = 0.02$ | 0.7388 (7) | 0.9650 (6) | 0.9990 (6) |
| $p = 0.05$ | 0.9674 (7) | 1.0000 (8) | 1.0000 (3) |
| Sig. level, $\alpha = 0.05$ | 0.2458 | 0.5064 | 0.7584 |
| Sig. level, $\alpha = 0.001$ | 0.0186 | 0.0443 | 0.0892 |
| **LR Test** | $k = 16$ | $k = 40$ | $k = 80$ |
| $p = 0.01$ | 0.6799 (10) | 0.9583 (10) | 0.9990 (10) |
| $p = 0.02$ | 0.8861 (10) | 0.9971 (10) | 1.0000 (8) |
| $p = 0.05$ | 0.9963 (10) | 1.0000 (7) | 1.0000 (3) |
| Sig. level, $\alpha = 0.05$ | 0.4428 | 0.7532 | 0.9491 |
| Sig. level, $\alpha = 0.001$ | 0.1781 | 0.4761 | 0.8075 |

the number of reads from individual lanes. The actual significance level of the LR test is relatively unaffected by the variance of the number of reads.

# 6 Conclusion

This paper has considered a statistical model for the detection of SNPs using DNA pooling. A simple test based on the maximum number of reads in a lane of a rarely observed allele is presented. On the basis of this test, we derive optimal pool sizes for the detection of rare alleles when the read rate is known. This test is compared to a likelihood ratio test, which is based on the number of reads of a rarely observed allele from each lane.

The actual significance level of these tests depends on how realistic the simple model presented in Section 2 is. However, more importantly, most deviations from this model lead to the test becoming more conservative, but have very little effect on the power of the test. It should be noted, however, that when the variance of the number of reads is large in comparison to the Poisson distribution, then the

Table 9: Effect of the overdispersion of the number of reads from a lane (pool sizes from 1 to 10, based on 10 000 simulations). In the case of no overdispersion the number of reads from a lane has a Poisson(20) distribution. The expected value and variance of the number of reads in the case of overdispersion are 20 and 80, respectively. Unless otherwise stated, the figures given are powers and optimal pool sizes (given in brackets) when $p = 0.02$. The significance levels are given for a nominal significance level of 5% (maximum significance levels are given for the LR test).

| **Maximum Test - No overdispersion** | $k = 16$ | $k = 40$ | $k = 80$ |
|---|---|---|---|
| $\varepsilon = 0.01$ | 0.6237(5) | 0.8624(3) | 0.9774(3) |
| $\varepsilon = 0.002$ | 0.7195(6) | 0.9603(6) | 0.9984(6) |
| Sig. level, $\varepsilon = 0.01$ | 0.0182 | 0.0449 | 0.0045 |
| Sig. level, $\varepsilon = 0.002$ | 0.0124 | 0.0307 | 0.0008 |
| **LR Test - No overdispersion** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.5738(4) | 0.8846(5) | 0.9889(5) |
| $\varepsilon = 0.002$ | 0.7446(9) | 0.9687(10) | 0.9993(10) |
| Sig. level, $\varepsilon = 0.01$ | 0.0111 | 0.0125 | 0.0121 |
| Sig. level, $\varepsilon = 0.002$ | 0.0097 | 0.0085 | 0.0040 |
| **Maximum Test - Overdispersion** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.5838(5) | 0.8150(3) | 0.9674(3) |
| $\varepsilon = 0.002$ | 0.6943(6) | 0.9487(7) | 0.9972(7) |
| Sig. level, $\varepsilon = 0.01$ | 0.0260 | 0.0641 | 0.0089 |
| Sig. level, $\varepsilon = 0.002$ | 0.0138 | 0.0352 | 0.0010 |
| **LR Test - Overdispersion** | $k = 16$ | $k = 40$ | $k = 80$ |
| $\varepsilon = 0.01$ | 0.5551(4) | 0.8743(5) | 0.9856(5) |
| $\varepsilon = 0.002$ | 0.7235(10) | 0.9645(10) | 0.9995(9) |
| Sig. level, $\varepsilon = 0.01$ | 0.0137 | 0.0106 | 0.0112 |
| Sig. level, $\varepsilon = 0.002$ | 0.0081 | 0.0088 | 0.0071 |

maximum test becomes less conservative and in some cases the actual significance level may exceed the nominal significance level.

The mathematical analysis presented generally assumes that the expected number of reads per lane is known and the probability of an error is a known constant. However, it was shown that the LR test can be adapted to practical cases, where an estimate of the probability of an error is given for each read. In addition,

the LR test does not use information regarding the expected number of reads. The maximum test can be adapted to the case where the expected read rate is unknown by calculating the critical value for the test based on the mean number of reads observed per lane. When the number of available lanes is small, this has a negative effect on the power of the maximum test. When the probability of an error is unknown, in order to use the maximum test an upper bound on the error rate should be used. This retains control of the probability of a type I error. In this case, the power of the test will be similar to the power obtained when the probability of error is equal to this upper bound.

In the case of Model 3 (when an error is made, each of the 3 possible nucleotides are equally likely), the *LR* test works better than the maximum test. Looking at these results more closely, it can be seen that the improvement is comparable to the improvement made under the maximum test when the error rate decreases by a factor of three. If it is felt that Model 3 (or a similar model) is a better description of reality than Model 1, then we can adapt the maximum test to obtain such a gain in power by taking the error rate to be the maximum of the probabilities of making each particular type of error.

For Models 1 and 2, the power of the maximum test is comparable to the power of the *LR* test. Using the asymptotically optimal pool size, the maximum test tends to be the more powerful one when the expected number of individuals with the minor allele is small (i.e., when the number of lanes and the minor allele frequency are small). As the expected number of individuals with the minor allele increases, a slight gain in power can be achieved using the *LR* test.

The simulations carried out in Section 5.4 give an indication of the practical use of pooling when the expected number of reads per lane is large (this could be the result of the PCR amplification of a particular section of the genome). Even with a relatively small number of available lanes, pooling can be used to detect minor alleles of frequency 2-5% with a high power. It is important to note that in such cases the power of detecting such alleles plateaus at pool sizes above the asymptotically optimal pool size. This means that when the cost of using extra lanes is greater than the cost of increasing the sample size via an increased pool size, one does not need a very accurate estimate of the optimal pool size in order to choose a pool size. One could use a pool size somewhat greater than the estimate of the asymptotically optimal pool size (i.e., such that the expected number of reads from an individual is equal to [or slightly smaller than] the critical value of the test based on an initial estimate of the read rate). The simulations also indicate that the theoretical estimate of power gives a reasonable estimate of the empirical power when such a pool size is used. It follows that although the concept of the asymptotically optimal pool size has its limitations (e.g. it requires knowledge of the read rate), it can be useful in estimating the power of detecting somewhat rare

alleles (frequency 2-5%) in such practical problems using both the maximum test and the LR test. Such estimation can also be adapted to estimating the power of the LR test in detecting such alleles over a wider range of problems (e.g. when the read rate is lower, but a larger number of lanes are available).

There are several ways in which the model presented here should be developed. Cutler and Jensen (2010) state that pooling leads to a loss of information, in particular on linkage disequilibrium. Barcoding (see Craig *et al.* , 2008, Kenny *et al.* , 2011) gives us data which are similar to the data obtained from sequencing individuals (albeit with a smaller number of reads per individual when pooling is used). It also preserves information on linkage disequilibrium. However, this comes at the cost of a more complex experimental procedure. Hence, an obvious adaptation of this model would take into account the possibility and costs of barcoding, together with the costs of using each lane. Also, other information may be useful, e.g. the cycle number of a site (see Druley *et al.* , 2009).

The model presented assumes that a) any read is equally likely to come from each of the individuals whose material is in a lane, b) the error rate of the sequencer is known and c) the number of reads comes from a Poisson distribution with known expected value. Cutler and Jensen (2010) argue that it is difficult to obtain equal concentrations of DNA from the individuals in a pool. However, the results of Kenny *et al.* (2011) using barcoding and human DNA from Chromosome 1 show that the assumption of equal amounts of probe material from each individual in a pool may be reasonable. Lynch (2008) argues that estimation of the error rate is difficult when it is relatively large compared to nucleotide diversity. However, as an increasing number of sequencing experiments have been carried out, estimates of the error rates using the Ewing and Green algorithm (see Ewing and Green , 1998) have become more accurate. The third assumption seems the most problematic. The expected number of reads may depend on the site (see Craig *et al.* , 2008, Palmieri and Schlötterer , 2009). Also, when pooling is used in combination with the PCR enrichment of chosen segments, the effect on the number of reads obtained is unpredictable (see Sham *et al.* (2002), Kenny *et al.* (2011)). However, suppose the read rate is large and the number of available lanes is small (which may well be the case in practical applications). Simulations show that the empirical power of the two tests is relatively unaffected by the pool size, as long as the pools are sufficiently large

The analysis presented here shows that the use of a simple threshold rule to detect SNPs is very efficient. Also, if we have information regarding the expected number of reads of a site from a lane, then optimal pool sizes can be derived for the detection of rare minor alleles. Further research should adapt this model to genetic barcoding and uncertainty regarding the expected number of reads.

# 7 Software

Software in the form of R code is available on request from the corresponding author (david.ramsey@ul.ie).

# References

Achaz G. (2008), Testing for neutrality in samples with sequencing errors. *Genetics* **179**, 1409–1424.

Balding D. J., Bishop M. and Cannings C. *eds.* (2008), *Handbook of statistical genetics*, 3rd edition. Hoboken, NJ: Wiley.

Benjamini Y. and Hochberg Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**(Suppl 1), 289–300.

Craig D. W., Pearson J. V., Szelinger S., Sekar A., Redman M., Corneveaux J. J., Pawlowski T. L., Laub T., Nunn G., Stephan D. A., Homer N. and Huentelman M. J. (2008), Identification of genetic variants using bar-coded multiplexed sequencing. *Nature Methods* **5** (Suppl 10), 887–893.

Cutler D. J. and Jensen J. D. (2010), To pool, or not to pool? *Genetics* **186**, 41–43.

Davison, A. (2008), *Statistical Models*. Cambridge: Cambridge University Press.

Druley T.E., Vallania F. L. M., Wegner D. J., Varley K. E., Knowles O. L., Bonds J. A., Robinson S. W., Doniger S. W., Hamvas A., Cole F.S., Fay J. C. and Mitra R.D. (2009), Quantification of rare allelic variants from pooled genomic DNA. *Nature Methods*, **6** (Suppl 4), 263–265.

Erlich Y., Chang K., Gordon A., Ronen R., Navon O., Rooks M., Hanon G. J. (2009), DNA sudoku-harnessing high-throughput sequencing for multiplexed specimen analysis. *Genome Research* **19**, 1243–1253.

Ewing B. and Green P. (1998), Base-calling of automated sequencer traces using *phred*. II. Error probabilities. *Genome Research* **8**, 186–194.

Futschik A. and Schlötterer C. (2010), Massively parallel sequencing of pooled DNA samples - the next generation of molecular markers. *Genetics* **186**, 207–218.

Holt K. E., Teo Y. Y., Li H., Nair S., Dougan G., Wain J. and Parkhill J. (2009), Detecting SNPs and estimating allele frequencies in clonal bacterial populations by sequencing pooled DNA. *Bioinformatics* **10**, 2074–2075.

Jiang R., Tavaré S. and Marjoram P. (2009), Population genetic inference from resequencing data. *Genetics* **181**, 187–197.

Kenny E. M., Cormican P., Gilks W. P., Gates A. S., O'Dushlaine C. T., Pinto C., Corvin A. P., Gill M., Morris D. W. (2010), Multiplex target enrichment using DNA indexing for ultra-high throughput SNP detection. *DNA Research* **18**(Supp 1), 31–38.

Knudsen B. and Miyamoto M. M. (2009), Accurate and fast methods to estimate the population mutation rate from error prone sequences. *BMC Bioinformatics* **10**, 247–258.

Lehmann E. (1986) *Testing Statistical Hypothesis*, 2nd edition New York: Springer.

Lynch M. (2008), Estimation of nucleotide diversity, disequilibrium coefficients and mutation rates from high-coverage genome sequencing projects. *Molecular Biology and Evolution* **25**(Suppl 11), 2409–2419.

Palmieri N. and Schlötterer C. (2009), Mapping accuracy of short reads from massively parallel sequencing and the implications for quantitative expression profiling. *PLoS One* **4**(Suppl 7), e6323.

Sham P., Bader J. S., Craig I., O'Donovan M. and Owen M. (2002), DNA pooling: A tool for large-scale association studies. *Nature Reviews Genetics* **3**, 862–871.